A2. Análisis y síntesis de voz

XUNTA DE GALICIA
CONSELLERÍA DE CULTURA, EDUCACIÓN E UNIVERSIDADE

IES de Teis
Avda. de Galicia, 101
36216 – Vigo
886 12 04 64
ies.teis@edu.xunta.es

GOBIERNO DE ESPAÑA
MINISTERIO DE EDUCACIÓN Y FORMACIÓN PROFESIONAL
Unión Europea-
NextGenerationEU

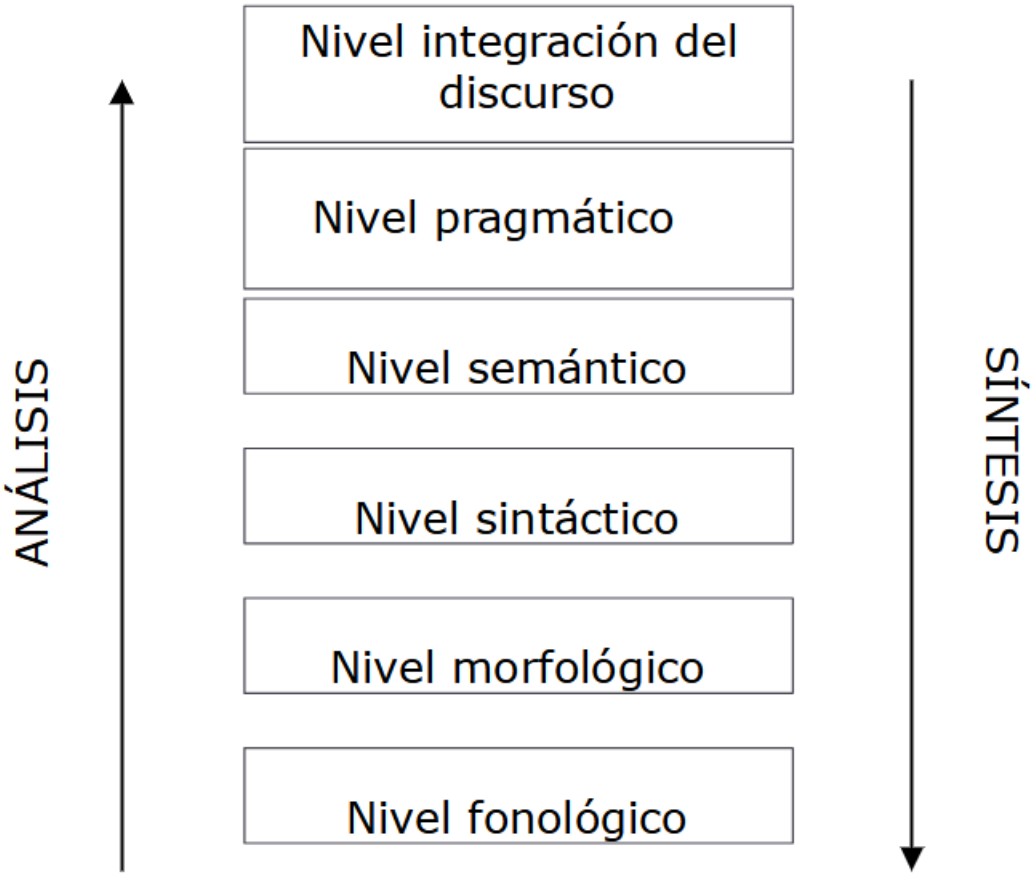
Índice.

A2. Análisis y síntesis de voz

1. Objetivo.

- Aprender cómo se forma la voz y sus características principales.
- Tener una idea básica de en qué se basa el reconocimiento del habla y de las técnicas empleadas.
- Conocer el estado actual de las tecnologías de interacción, como la síntesis o el reconocimiento de voz.
- Ver aplicaciones actuales de esta tecnología.

2. Arquitectura de niveles.



2. Arquitectura de niveles.

Los niveles son los siguientes:

- **Nivel fonológico** → conversión de voz a texto bajo los requisitos del conocimientos de los fonemas del lenguaje y un algoritmo de reconocimiento.

  Resulta muy importante el tratamiento de la ambigüedad → requiere un conocimiento de los niveles superiores (al menos morfológico y sintáctico) con el fin de evitar la confusión del significado y la consiguiente pérdida de información.

2. Arquitectura de niveles.

- **Nivel morfológico** → análisis de lexemas, categoría gramatical, atributos propios de categoría. Bajo requisitos de conocimiento de los formantes (raíz y desinencias.) y gramática de las palabras.

  Una lista de palabras no suele valer SINO una base léxica, como almacén de información fundamentalmente morfológica, aprovechando las regularidades de la lengua y escrita para lingüistas; sin embargo, no se debe sobregenerar ni sobreaceptar.

Ambigüedad: suelo

2. Arquitectura de niveles.

- **Nivel sintáctico** → estructura en árbol de agrupación de palabras y relaciones. A partir de información morfológica de palabras (léxico) y de la gramática de la frase.

  Una gramática general es difícil → por no decir que imposible.

  Complejidad del léxico vs complejidad de la gramática (directamente proporcional).

  Ambigüedad:

  Se tomó el helado con cuchara
  Se tomó el helado con vainilla

2. Arquitectura de niveles.

- **Nivel semántico** → significado literal de la frase, en función del mundo y de las reglas semánticas. Totalmente dependiente de la aplicación concreta (dominio restringido).

Ambigüedad → “Pasé delante del banco”.

2. Arquitectura de niveles.

- **Nivel pragmático** →significado literal de frase vs. Significado real de frase

2. Arquitectura de niveles.

- **Nivel de integración del discurso** → significado de la frase aislada vs significado en contexto

3. Ambigüedades.

En la mayoría de los casos, la resolución de una ambigüedad de un nivel requiere análisis de los niveles superiores.

Algunas de las dificultades que nos podemos encontrar en el análisis son las siguientes:

- Modelos lingüísticos insuficientes.
- La sintaxis implica gramática dependiente del contexto.
- Tratamiento de la semántica.
- Niveles superiores a la semántica aún más complejos.
- Abordable sólo parcialmente con arquitectura de niveles.
- Aplicaciones muy variadas → solución general difícil.
- Diferencias entre lenguas.
- Inserción de conocimiento manual.

3. Ambigüedades.

La solución a estas ambigüedades suele producirse solucionando previamente subproblemas pequeños.

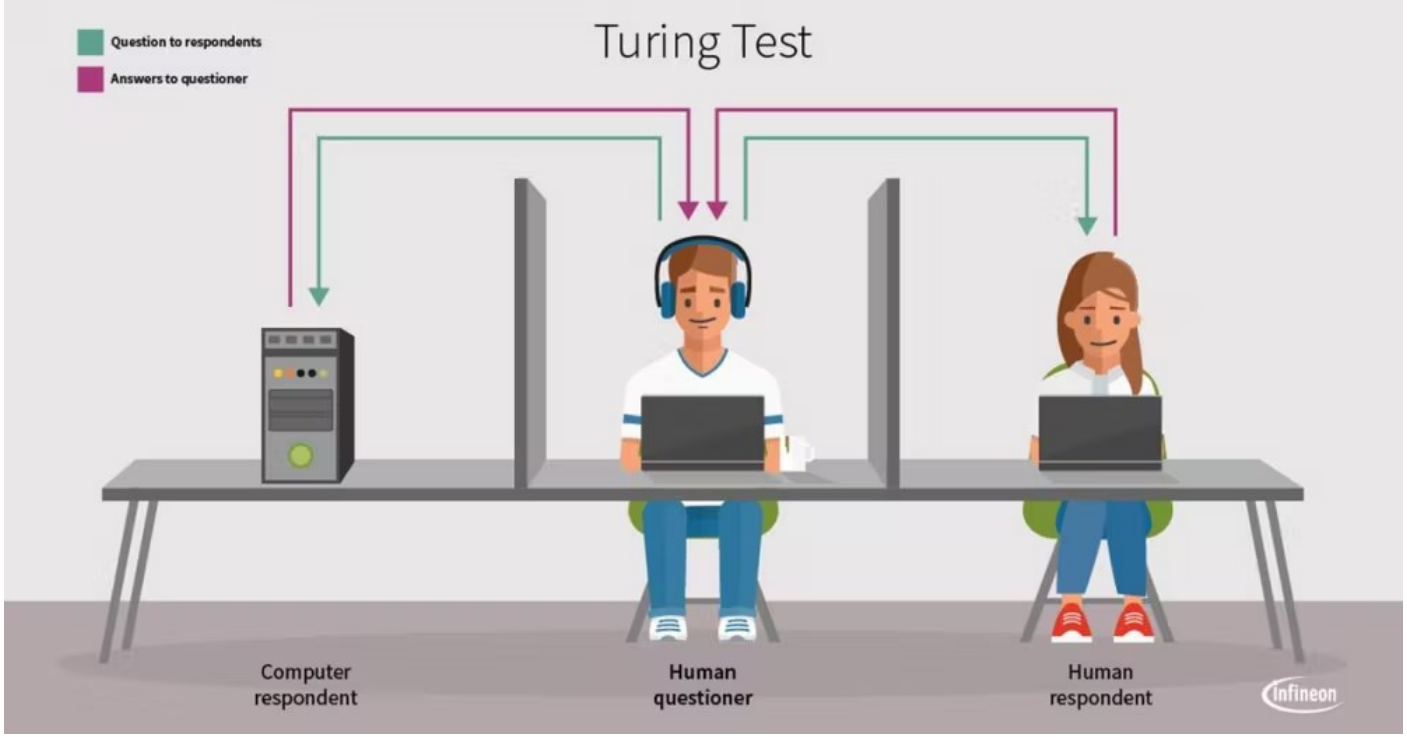
Así, es posible desarrollar sistemas realmente útiles.

El tiempo corre a nuestro favor:

- Ordenadores más potentes.
- Formalismos más desarrollados.
- Más experiencias y desarrollos.

4. Análisis de la voz.

En 1950 un matemático británico, llamado Alan Turing, propuso comparar al Hombre con la máquina, proponiendo a una persona para que mantuviera una conversación de texto a ciegas con un humano y un ordenador: esta persona debía averiguar en 5 minutos cuál de sus interlocutores era el ordenador.



4. Análisis de la voz.

El año 2016 marcó un hito en la tecnología, porque los investigadores de Microsoft llevaron el reconocimiento de voz a un nuevo nivel, desarrollando una IA capaz de realizar transcripciones de audio a un nivel equivalente al de un humano.

En 2017 la IA llegó a superar la competencia humana, demostrando que la máquina entiende cada vez mejor el lenguaje natural.

4.1. Funcionamiento de la IA de voz.

La **ASR** (Automatic Speech Recognition) es un software de reconocimiento del habla que permite al usuario emitir una solicitud mediante voz, que posteriormente se convierte en texto tras analizar el contexto.

El análisis es todo un reto debido a que la pronunciación de los fonemas puede tener varios sentidos, y es en este momento en el que la IA entra en juego para encontrar el sentido real de la pregunta emitida.

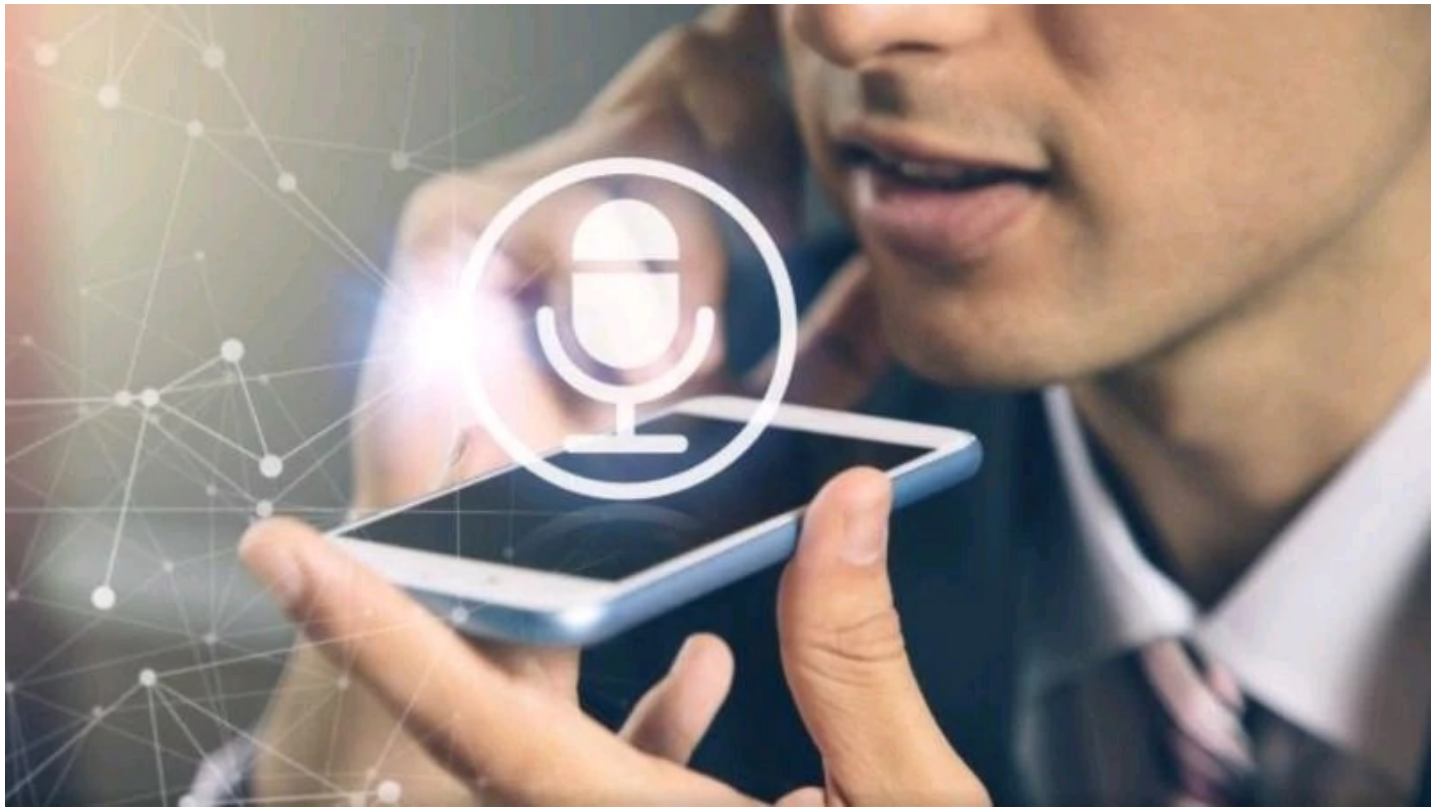
La IA **paralingüística** consiste en entrenar a los algoritmos para detectar las características inherentes al locutor sin relación con el contenido verbal pronunciado

A2. Análisis y síntesis de voz

4.1. Funcionamiento de la IA de voz.

El funcionamiento del reconocimiento de voz desde un punto de vista técnico es el siguiente:

- Un micrófono capta la voz de una persona.
- La voz, en forma de vibración, se transforma en señal eléctrica ondulatoria.
- El software de reconocimiento de voz la analiza en forma de fonemas y reconstruye palabras.



4.2. Tecnologías de análisis de voz.

La forma en que nos comunicamos con la tecnología ha evolucionado muchísimo en los últimos años → desde los teclados, las pantallas táctiles hasta llegar a la voz.

La **interacción a través de la voz** se ha ido introduciendo poco a poco en la vida de los usuarios, especialmente tras la llegada de los asistentes de voz y las nuevas soluciones en el ámbito telefónico para:

- Facilitar la redirección de llamadas.
- Facilitar la recopilación de información básica de quien llama.
- Agilizar gestiones como reservas de cine o teatro.

4.2. Tecnologías de análisis de voz.

El sector ha evolucionado a un ritmo vertiginoso e integrando las nuevas herramientas para mejorar la calidad y la eficiencia.

Las tecnologías de voz que supondrán una revolución en la forma en que nos comunicamos son:

- **Reconocimiento automático del habla (ASR, Automatic Speech Recognition)** → responsable del entendimiento y transcripción del DNI o teléfono de una persona a un operador virtual, evitando tener que marcar vía teclado y crear una interacción más sencilla con el usuario.
- **Análisis del habla** → tecnología capaz de analizar las llamadas de voz en tiempo real para detectar emociones, tonos o, incluso, estrés. De esta forma, las compañías pueden mejorar servicios y adaptarse a las necesidades de su público para mejorar la forma de abordar problemas.
- **Voice Engine Optimization (VEO)** → los altavoces inteligentes con asistente de voz son ya cada día un elemento más en los hogares, y la interacción con ellos se ha convertido en algo cotidiano, incrementándose las búsquedas mediante voz.
- **Biometría de voz** → tecnología encargada del reconocimiento por medio de la huella vocal de cada individuo, en lugar del uso de huella dactilar o rasgos físicos.

4.3. Evolución de las interfaces hombre-máquina (HMI).

Nuestra relación con las máquinas se realiza a través de interfaces que han ido evolucionando muchísimo a través de las décadas.

El origen de dicha interfaz es el teclado, y se lo debemos a Eliphalet Remington que lo inventó en 1714.

En 1930, al necesitarse una interfaz para el fax, se adoptó su teclado QWERTY como entrada de información.

En 1948 apareció el primer teclado electrónico con los ordenadores Binac.

4.3. Evolución de las interfaces hombre-máquina (HMI).

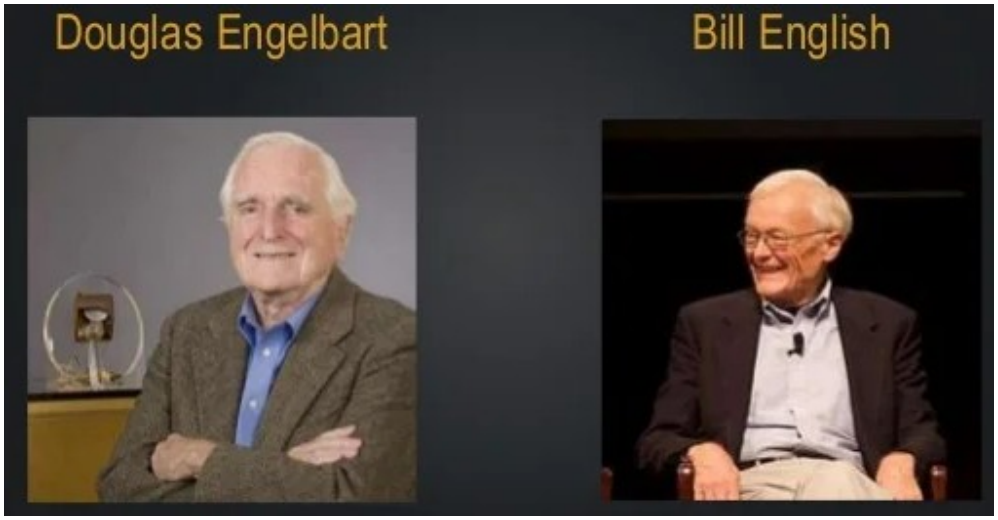
En 1945 Vannevar Bush (MIT) publicó un artículo llamado ‘As we may think’, en el que hablaba de Memex, su ideación de cómo sería un interfaz visual con pantallas que seguía hipervínculos.

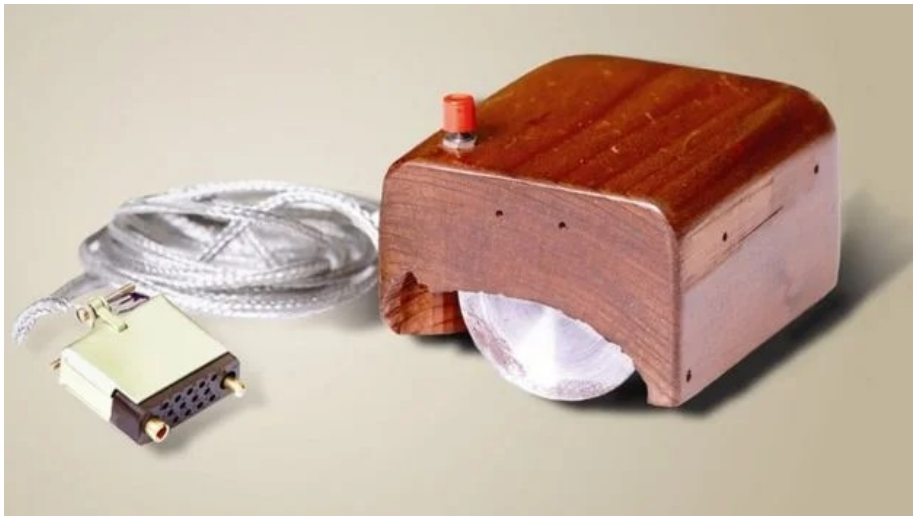
Aunque usaba la tecnología de la época, poco se alejó de lo que sería el primer ordenador.

4.3. Evolución de las interfaces hombre-máquina (HMI).

En 1960 Douglas Engelbart y Bill English trabajaban en conseguir algo mejor que sólo una pantalla estática manejada por comandos, e inventaron algo de uso muy corriente hoy en día: el ratón de ordenador.

Tras el ratón aparecieron entornos más amigables y las interfaces gráficas de usuario.





4.3. Evolución de las interfaces hombre-máquina (HMI).

Una de las últimas interfaces de comunicación hombre-máquina es, precisamente, el control por voz, que brinda la oportunidad de comunicarnos con las máquinas a través de nuestro propio lenguaje.

5. Síntesis de la voz.

La síntesis de voz es el proceso de replicación de la comunicación verbal a través de un dispositivo artificial.

La primera máquina parlante fue creada por Wolfgang von Kempelen en 1700.

El habla se producía a través de un fuelle de cocina (actuaba como pulmón), una caña de gaita (simulaba la glotis) y una campana de clarinete (servía de boca).

5.1. Características de un sintetizador de voz.

Las características que debe tener un sintetizador de voz son:

- Inteligibilidad → relacionada con la facilidad para comprender la señal oral.
- Calidad → indicador de la naturalidad de los sonidos.

5.2. Tipos de sintetizadores de voz.

Un sintetizador de voz puede ser de alguno de los siguientes tipos:

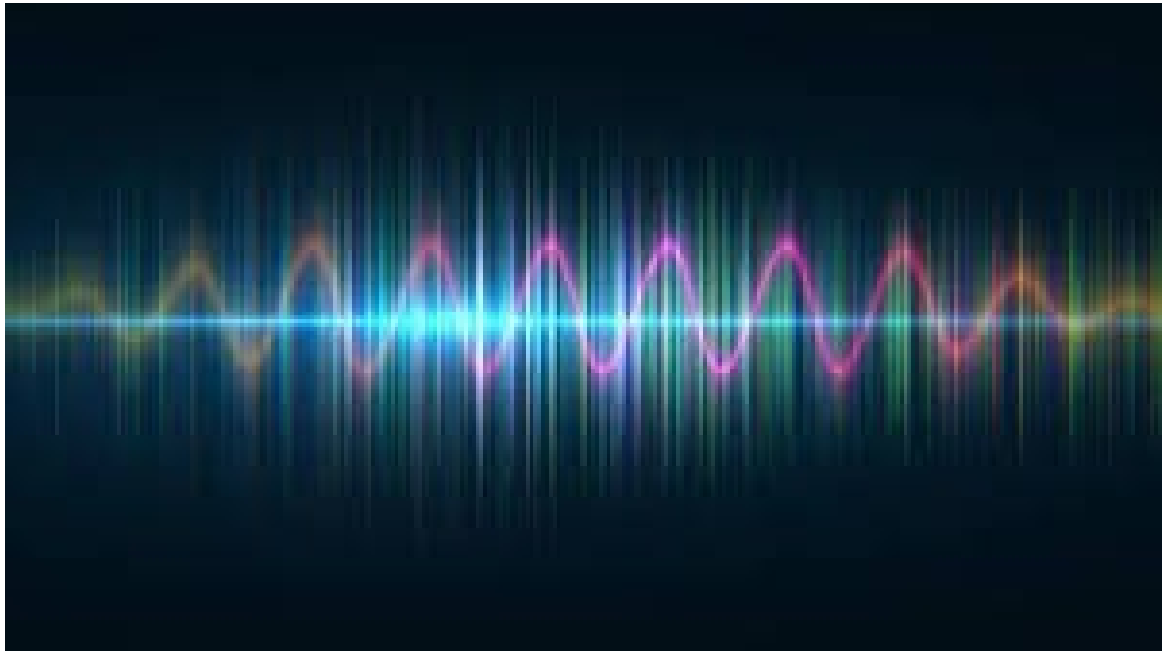
- **Sistema de respuesta oral** → basados en la reproducción de segmentos de voz previamente grabados. Ejemplo es el caso de información telefónica.

- **Convertidor texto-voz** → sistemas capaces de convertir cualquier cadena de texto de entrada en una señal de voz.

5.3. Ventajas y desventajas.

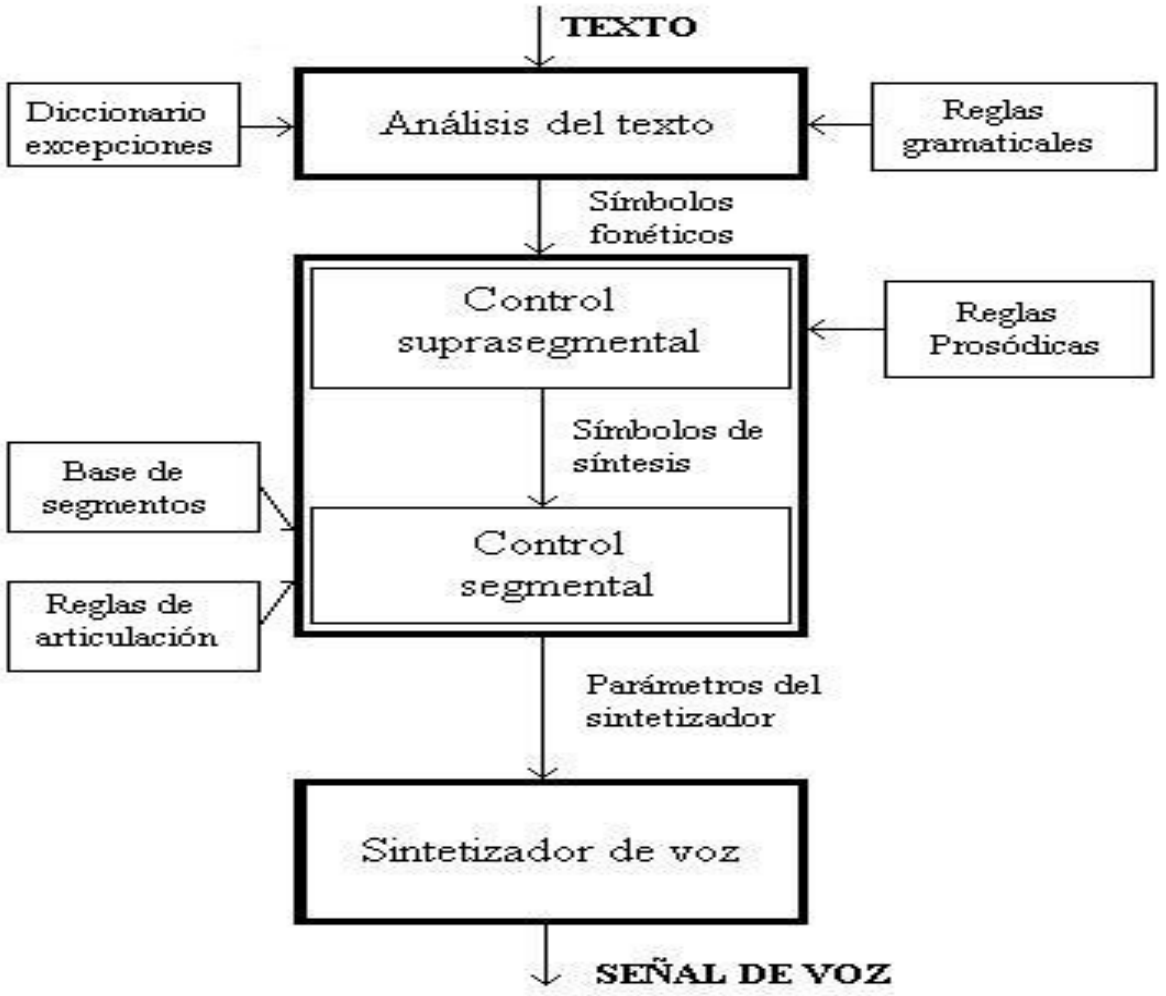
Las ventajas y desventajas de los sistemas de respuesta oral frente a los convertidores texto-voz se pueden afrontar desde las siguientes perspectivas:

- **Número de frases sintetizadas** → los sistemas de respuesta oral sólo permiten sintetizar un número muy limitado frente a cualquier frase de entrada de los convertidores texto-voz.
- **Complejidad** → los convertidores texto-voz son mucho más complejos.
- **Flexibilidad** → los convertidores texto-voz son mucho más flexibles.
- **Gasto de memoria** → los sistemas de respuesta oral requieren un menor consumo de memoria.



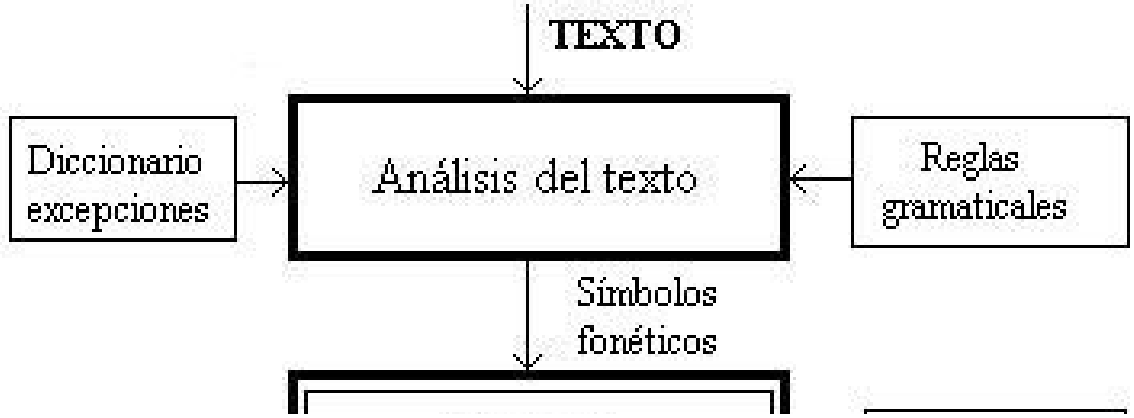
5.4. Esquema de un sintetizador: Conversor texto-voz.

El esquema del sintetizador es el siguiente:



5.4. Esquema de un sintetizador: Conversor texto-voz.

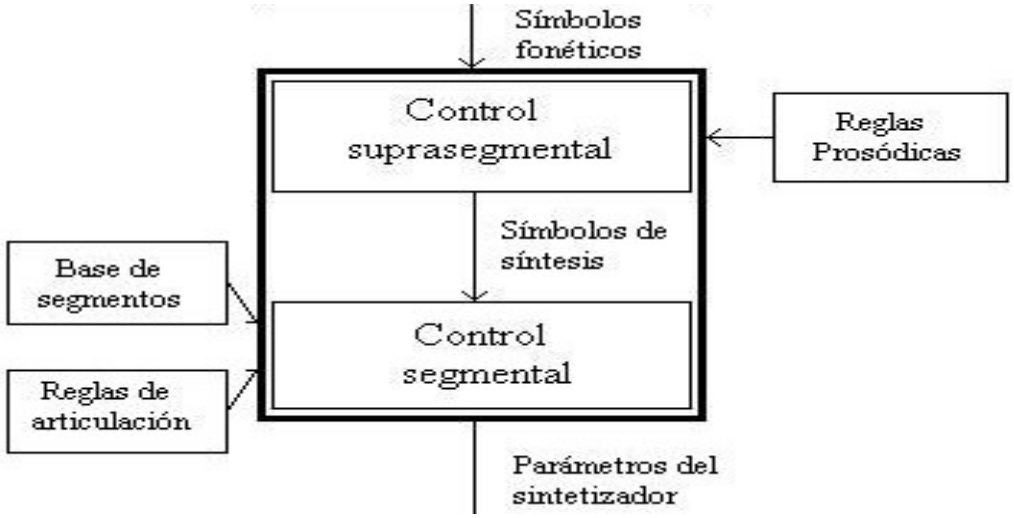
1ª Etapa: Análisis del texto



Función → realiza la conversión de los símbolos fonéticos que integran el texto escrito, usando las reglas gramaticales propia del lenguaje.

5.4. Esquema de un sintetizador: Conversor texto-voz.

2ª Etapa: Generación de prosodia



Función → se divide en dos bloques:

- Control suprasegmental → trata la entonación de la frase en su conjunto.
- Control segmental → control la micromelodía, o los fenómenos locales de coarticulación, acentuación, ...

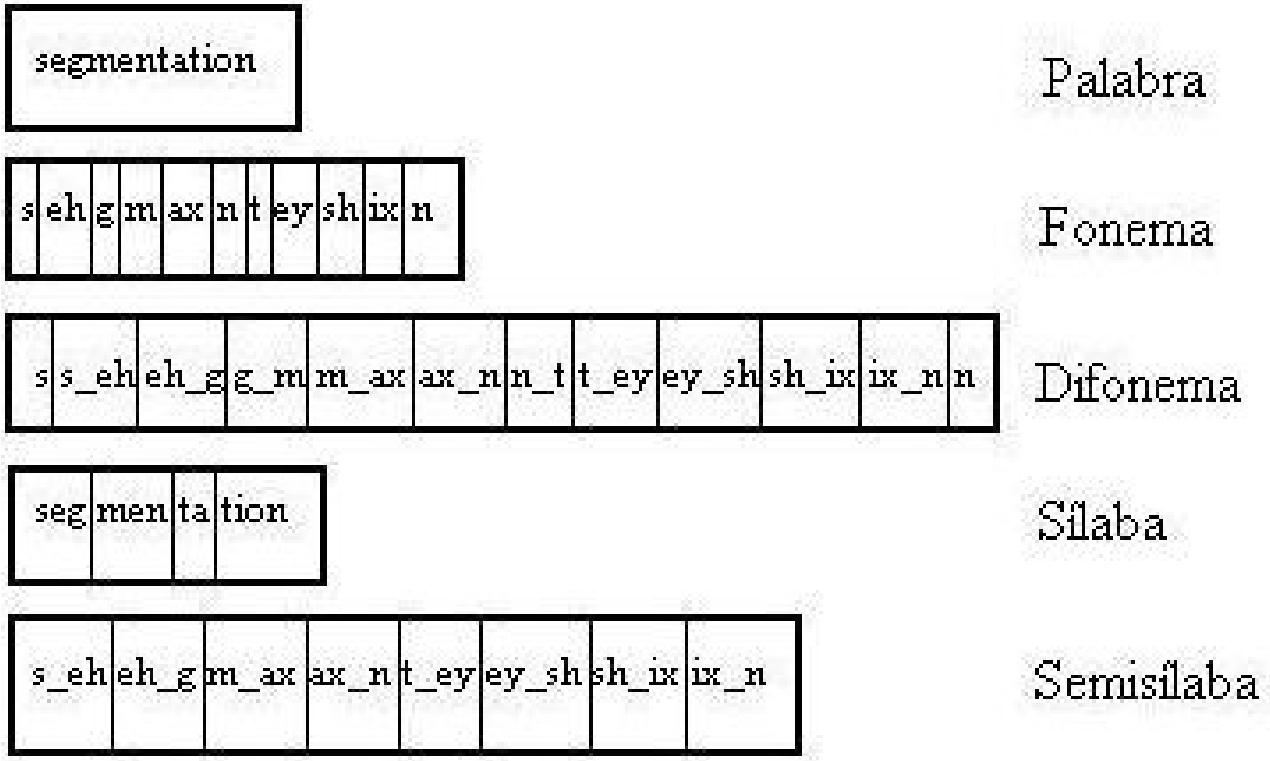
5.4. Esquema de un sintetizador: Conversor texto-voz.

3ª Etapa: Síntesis de la voz

Función → genera la seña acústica a partir de los parámetros extraídos en los bloques anteriores.

5.5. Ejemplo de división de unidades lingüísticas.

Un ejemplo de unidades lingüísticas son las siguientes:



Ejemplo de división en unidades

5.6. Desafíos involucrados en la síntesis del habla.

Los desafíos son los siguientes:

- Acomodación de palabras pronunciadas de forma distinta que tienen la misma ortografía, según el contexto.
- Inferencia de cómo expandir un ‘no.’ basado en la palabra, el número y la puntuación circundantes. → 1465 como ‘mil cuatrocientros sesenta y cinco’, ‘catorce sesenta y cinco’, `’catorcecientos sesenta y cinco’.
- Ambigüedad en abreviaturas → ‘en’ para pulgadas, contra la palabra ‘en’.
- El enfoque basado en el diccionario del proceso de texto a fonema falla completamente para cualquier palabra que se pueda encontrar en el diccionario.
- El enfoque basado en reglas del proceso de texto a fonema falla porque el esquema tiene en cuenta ortografías o pronunciaciones inusuales debido al aumento considerable de la sofisticación de las reglas.
- Dificultad en la evaluación confiable de los sistemas de síntesis de voz debido a la falta de estándares de desempeño aceptados.
- Desplazamiento del contorno tonal de la oración, según sea una expresión afirmativa, interrogativa o exclamativa.

6. Aplicaciones.

Hay cada vez un sinnúmero de aplicaciones, pero a febrero de 2022, podemos decir que las 10 mejores soluciones de texto a voz de uso comercial y personal son las siguientes:

- **Murf.ai** (https://murf.ai/?lmref=oYsoew) → proporciona un generador de voz de IA versátil, con más de 100 voces de texto a voz realista, en más de 15 idiomas.
- **TTSReader** (https://ttsreader.com/) → cuadro de texto en el que se puede escribir o pegar cualquier texto y hacer clic en el botón de reproducción. Admite muchos idiomas, acentos y variaciones de velocidad.
- **Wideo** (https://wideo.co/text-to-speech/?utm_source=cjaffilate&utm_medium=affiliate&utm_campaign=cjcampaign&cjevent=a3195e27a97d11ec8123035b0a180513) → forma sencilla y rápida de convertir texto en voz. Se escribe texto en el cuadro, se elige entre las voces y la velocidad.
- **NaturalReader** (https://www.naturalreaders.com/) → poderosa conversión de texto a voz, con lectura clara y de alta calidad con voces de sonido natural.
- **ReadSpeaker** (https://www.readspeaker.com/) → permite introducir productos en el mercado con soluciones de voz, a través de elección de idioma y voz.
- **Notevibes** (https://notevibes.com/) → conversión en línea de texto a 201 voces con sonido natural.
- **FreeTTS** (https://freetts.com/) → solución gratuita para convertir entre más de 35 idiomas, pudiendo elegir el tipo de voz.
- **Google Cloud** (https://cloud.google.com/text-to-speech) → funciona con tecnologías de IA de Google, ayudando a mejorar las interacciones con los clientes a través de respuestas inteligentes y realistas.
- **Watson** (https://www.ibm.com/cloud/watson-text-to-speech) → convierte texto en discurso de voz con sonido natural en varios idiomas.
- **Amazon Polly** (https://aws.amazon.com/es/polly/) → forma eficaz de conversión de texto en habla humana, utilizando el aprendizaje profundo para sintetizar un habla que suene natural.